Cluster Computing (2019) 22:S4277–S4282
https://doi.org/10.1007/s10586-018-1861-4

Method and application of data defect analysis based on linear discriminant regression of far subspace

Wendai Lv[1]

Received: 19 December 2017 / Revised: 10 January 2018 / Accepted: 14 January 2018 / Published online: 29 January 2018

Abstract
In order to improve effectiveness of analysis algorithm for internal control deficiency in familiar listed companies, an analysis method of internal control deficiency in familiar listed companies based on linear discriminant regression of far subspace was proposed. Firstly, influencing effect model of internal control deficiency disclosure on corporate executive compensation level and compensation-performance sensitivity was built and analyzed; secondly, closest subspace classifier was used to measure the relationship between internal control deficiency of familiar listed companies and single class based on linear discriminant regression, and algorithm development based on farthest distance between source vector and projection vector was implemented; finally, effectiveness of proposed methods was verified by empirical analysis.

Keywords Far subspace · Linear discriminant regression · Familiar listed company · Internal control deficiency

1 Introduction

Basic norms of enterprise internal control requires that familiar listed companies evaluate corporate internal control status and disclose evaluation report [1]. We can see from report disclosed by familiar listed company that management has great decision-making power on problems such as whether it is necessary to disclose existing internal control deficiency in company, use what kind of forms to disclose and disclosure extent and so on based on different intentions. Critical role of internal control is to maintain reliability of accounting information and prevent great investment losses [2]. Operation status of corporate internal control influences corporate investment, financing and other great decisions directly. Corporate external stakeholders (including investors, creditors, potential investors and social public) have paid attention to construction, system improvement, operation mechanism, operation status and influential effect and so on of corporate internal control system and strive to master information to make decisions further.

✉ Wendai Lv
quyahhb211@163.com

[1] Department of Accounting, Faculty of Management and Economics, Kunming University of Science and Technology, Kunming 650000, Yunnan, China

Public Company Accounting Oversight Board (PCAOB for short) indicated in 2004 that when internal control of financial reports has substantial deficiency, the possibility that corporate financial reports have material misstatement is higher. Under the condition of weaker internal control system, the company may have problems. For example, the system is imperfect; staff lack diligence, and cost benefits cannot be calculated accurately and so on, which results in that quality of corporate financial reports reduces and financial information is unreliable. Internal control is the first "defense line" to prevent material misstatement of corporate financial report, while auditor is the other "defense line" to reduce misstatement based on investor protection [3]. In order to allocate resource effectively, auditors allocate more resources to areas may have risk of material misstatement based on risk-oriented audit standard. Therefore, if auditors pass on higher costs to customers, audit fees will be improved along with increase of internal risk. Audit test cannot find all possible misstatement (such as fraudulent conduct of collusive management), so it is necessary for auditors to implement risk pricing and improve rate for companies with higher risk of audit service to compensate increased litigation and insurance fees and construct enough financial reserves [4–6].

Many researches have proved the relationship of corporate internal control deficiency, income quality reduction,

earnings management and risk increase, and proved that internal control deficiency represents risk of users of financial reports. Do these previous research results mean that internal control deficiency has an effect on audit fee? Whether existing internal control deficiency of familiar listed companies influences audit fees is discussed from the view of cross section and internal control deficiency is classified specifically to study differences of impacts of internal control deficiency with different classification on audit fees in the Thesis by using data of A-share familiar listed company in 2009 and 2010 in China; at the same time, the impacts on audit fees after the Company corrects internal control deficiencies are studied from inter-temporal view.

Aimed at internal control deficiency of familiar listed enterprises, research was implemented in the Thesis. The relationship between internal control deficiency of familiar listed companies and single class was measured by closest subspace classifier based on linear discriminant regression, and algorithm development was implemented based on farthest distance between source vector and projection vector to realize effectiveness of corporate internal control deficiency. Empirical analysis result verified effectiveness of algorithm [7].

2 Research design

2.1 Sample source and processing

Corporate internal control guidance was implemented in familiar listed companies of Shanghai and Shenzhen main board on January 1, 2012. Therefore, familiar listed companies of A-share main board in Shanghai and Shenzhen stock markets in 2012 are selected to be research samples in the Thesis. As financial characteristics of financial industry and ST Company are relatively special, the above samples are deleted in the research of the Thesis and companies with incomplete data are also deleted at the same time. Annual data of 809 companies are gotten in total finally. We arranged internal control audit report, internal control evaluation report and corporate annual report, and collected internal control deficiency and severity division, internal control audit opinions and other related data manually. Other financial data are derived from database of CSMAR Chinese stock research market and CCER Chinese economy and finance. In order to guarantee data accuracy, we drew part samples randomly to contrast with related announcement again after data integration, verified data and guaranteed data reality in the Thesis. In order to avoid occurrence of abnormal value, continuous variables in the model are provided with truncation (winSORiZE) of up and down 1% quantile.

2.2 Model building and research variable

Taking existing research results as reference, we built the following model to explore influential effect model of internal control deficiency disclosure on corporate executive compensation level and compensation-performance sensitivity respectively:

$$\begin{aligned} Ln\,(Comp_i) &= \beta_0 + \beta_1 ICW_i + \beta_2 ROA_i \\ &+\beta_3 ICW_i * BOA_i + \beta_4 SIZE_i + \beta_5 LEV_i + \beta_6 SEG_i \\ &+\beta_7 SOE_i + \beta_8 OWN_i + \beta_9 NAUDIT_i + \beta_{10} BIG4_i \\ &+\beta_{11} INDI_i + \beta_{12} LITI_i + \beta_{13} \sum Industry_i + \varepsilon_i \end{aligned} \quad (1)$$

$$\begin{aligned} Ln\,(CFO_i) &= \beta_0 + \beta_1 ICW_i + \beta_2 ROA_i \\ &+\beta_3 ICW_i * BOA_i + \beta_4 SIZE_i + \beta_5 LEV_i + \beta_6 SEG_i \\ &+\beta_7 SOE_i + \beta_8 OWN_i + \beta_9 NAUDIT_i + \beta_{10} BIG4_i \\ &+\beta_{11} INDI_i + \beta_{12} LITI_i \\ &+\beta_{13} \sum Industry_i + \varepsilon_i \end{aligned} \quad (2)$$

where $Comp_i$ is annual salary of the top three executives in the Company i. Researches indicated that when internal control information is disclosed compulsorily, CFO in corporate executives undertakes more risks, so natural logarithm of current compensation CFO_i of Finance Director in the Thesis is as the second substitutive variable of executive compensation. ROA_i is return on total assets of i company in 2012 to measure corporate performance of company i. According to definition and classification of Chinese corporate internal control deficiency in *evaluation guidance* of *corporate internal control guidance*, that familiar listed companies disclose their internal control deficiency (ICW) is regarded as core dependent variable of research in the Thesis: when familiar listed companies disclose that its internal control has deficiencies, $ICW_i = 1$; otherwise, $ICW_i = 0$. Association relationship of ICW_i, $ICW_i * BOA_i$ and corporate executive compensation is mainly explored in the Thesis.

In addition, we controlled the following variables: (1) corporate financial characteristics. $SIZE_i$ is natural logarithm of total assets of company i; LEV_i financial leverage is corporate asset-liability ratio. (2) Corporate organization characteristics. SEG_i is natural logarithm of subsidiaries no. of company; SOE_i is the property of corporate actual controller, when corporate actual controller is state-owned, $SOE_i = 1$, otherwise, $SOE_i = 0$; OWN_i measures ownership concentration of company i and is shareholding ratio of current top 10 shareholders in company i. (3) Corporate supervision characteristics. $NAUDIT_i$ is internal control audit opinions in current year of company i. When standardized and unqualified audit opinions are issued for internal control, $NAUDIT_i = 1$, otherwise, $NAUDIT_i = 0$; $BIG4_i$ is whether the accounting firm employed by company i in current year is "four big", when company i employs "four big" accounting firms as auditors in the current year,

$BIG4_i = 1$, otherwise, $BIG4_i = 0$; $INDI_i$ is log value of times for which independent directors participate in General Meeting of Shareholders and is substitute variable to measure whether dependent directors their responsibilities; $LITI_i$ is whether company i is confronted with lawsuits in the current year, when company i discloses that it is confronted with lawsuits in the current year, $LITI_i = 1$, otherwise, $LITI_i = 0$.

3 Linear discriminant analysis

3.1 Algorithm description

Assuming that there are N objects, including p_i training data $i = 1, 2, \ldots, N$ from i classification, project data to subspace; matrix W includes all feature vectors from N objects, namely $W = [W_1, \ldots, W_i, \ldots, W_N]$. In order to apply specific models of regression analysis estimation class, group column vectors of members of related class $w_{i,j}$ into collection. Therefore, as for class i, there is:

$$W_i = \left[w_{i,1}, \ldots, w_{i,j}, \ldots, w_{i,pi}\right] \in R^{L \times pi} \tag{3}$$

In the above equation, each vector is a column vector with size of $L \times 1$. Class i is expressed by vector space W_i during training stage. It is called predictor of each object.

If y belongs to class i, it can be expressed by a linear combination of training data in class i. The definition is as follows:

$$y = W_i \beta_i + e, i = 1, 2, \ldots, N \tag{4}$$

In the above equation, $\beta_i \in R^{p_i \times l}$ is vector of regression parameter, $\tilde{\beta}_i$ is an error vector, and mean value of its independent and identically distributed random variables is zero. Objective of linear regression is to find $\tilde{\beta}_i$ to minimize residual error.

It can be known from literature that linear regression is developed based on minimum distance between source vector and projection vector. If source sector belongs to subspace of class i, prediction vector $\tilde{y}_i$ will be the vector that is closest to source vector. Calculating Euclidean distance between source vector and predictive response vector can confirm identity i^* , namely:

$$i^* = \arg\min_i \|\tilde{y}_i - y\|, \quad i = 1, 2, \ldots, N \tag{5}$$

3.2 Far subspace classifier

Sample data of known specified object class are located in a class-specific subspace, and then it can be assumed that each class-specific subspace has some unique basic vectors. These vectors are mutually independent with vectors of other classes. Therefore, an innovative farthest recognition idea is proposed in the Thesis. When a query datum is expressed by all prototype data including prototype data in the class of this query data, if build a "leave-one-class-out" subspace for class i with all training data including data in class i, the distance between sample of class i and the subspace is maximum in all "leave-one-class-out" subspaces. Because only "leave-one-class-out" subspace in this class is exclusive in unique basic vectors of class i.

Based on the above analysis, far subspace (FS) classifier is proposed in the Thesis. Firstly, linear model of "leave-one-class-out" subspace in class i is built by arranging $N - 1$ class-specific subspace model:

$$B_i = \left[A_1 A_2 \ldots A_{i-1} A_{i+1} \ldots A_N\right] \tag{6}$$

Being similar to NS classifier, FS classifier is still a simple method based on linear discriminant analysis. Certain definition of FS variable is distance from query data to "leave-one-class-out" subspace. Calculation is as follows:

$$l_i = \|y - \bar{y}_i\|_2 \tag{7}$$

In the above equation, $\bar{y}_i = B_i \left(B_i^T B_i\right)^{-1} B_i^T y$, make $B_i^T B_i$ reversible by PCA and remove small principal component. Finally classify probe to the class with maximum distance l_i. FS classification method is as Algorithm 1.

Algorithm 1 Far subspace classification

1: Input: class-specific model $A_i \in R^{q \times pi}$, "leave-one-class-out" model $B_i = \left[A_1 A_2 \ldots A_{i-1} A_{i+1} \ldots A_N\right]$, $i = 1, 2, \ldots, N$ and test data vector $y \in R^{q \times 1}$;
2: Calculate projection y on subspace A_i, and $\hat{y}_i = A_i (A_i^T A_i)^{-1} A_i^T y$;
3: Measure distance between y and $\hat{y}_i$, and $d_i = \left\|y - \hat{y}_i\right\|_2$;
4: Calculate projection y on each "leave-one-class-out" subspace B_i, and $\bar{y}_i = B_i (B_i^T B_i)^{-1} B_i^T y$;
5: Measure distance between y and $\bar{y}_i$, and $l_i = \|y = \bar{y}_i\|_2$;
6: Calculate decision variable $j_i = d_i / l_i$, and $i = 1, 2, \ldots, N$;
7: Output: test data from class with minimum distance j_i .

Then, research of NFS and FS classifier will be explained in the Thesis, and whether proposed NFS classification method can be regarded as a single classification rule to substitute simple hybrid method will be discussed. It is known that some class-specific farthest data are usually located in a linear subspace, and each class-specific subspace has some unique basic vector, namely class-specific subspace is not only related to common basic vectors shared by many class-specific subspaces, but also has its own basic vectors. When prototype data is more enough or class-specific subspace

is realized, query data can use class-specific model of its own class to express, because their classes do not have these unique basic vectors. Namely NS classifier can make full use of unique basic vectors for classification.

Ideal class-specific subspace shall cover large range change of farthest data, but farthest identification usually has small sample (SSS) problems. Actual class-specific subspace only can be constituted by prototype data with fewer amounts, and the amount is much fewer than that of ideal class-specific subspace. Namely SSS problem usually leads in imperfection of class-specific subspace, which may leave out some commonly-used basic vectors. Therefore, although data from the same class are located in a class-specific subspace, distance between measured test data and imperfect class-specific subspace will be deviated from actual distance value.

4 Empirical test and result analysis

Empirical result for influential effect of internal control deficiency for complete sample on executive compensation is reported in Table 1. VIF values of models are less than 3 through test, which indicates that serious multicolinearity problem does not exist in all vectors of models. From the result, executive compensation of familiar listed company in China ($Comp_i$ and CFO_i) is significantly and positively related to corporate performance (ROA_i) at the level of 10%. It indicates that executive compensation of familiar listed companies in China has significant performance sensibility. It is interesting that we found that executive compensation is significantly and positively related to variable ICW_i, namely executive compensation level of those familiar listed companies that disclose their internal control have deficiencies is higher. At the same time, we noticed that coefficient before $ICW_i * BOA_i$ is significantly positive and greater than coefficient before ROA_i, namely when company discloses internal control deficiencies, performance sensibility of executive compensation in familiar listed company is improved. We refine research objects into CFO in familiar listed company further. We can see from empirical result that compensation level of CFO is still significantly and positively related to variable ICW_i, namely CFO in those familiar listed companies that disclose internal control deficiency has higher compensation level. At the same time, being similar with result of executive compensation, disclosing internal control deficiency improves performance sensibility of compensation of CFO.

Table 1 Internal control deficiency and executive compensation sensibility

Complete sample ($N = 809$)	$Ln\,(Comp_i)$ coefficient	$Ln\,(CFO_i)$ coefficient
(Intercept)	8.956 (23.410)	8.224 (19.727)
ICW_i	0.237 (3.819)	0.120 (2.057)
ROA_i	0.411 (1.774)	0.278 (1.715)
$ICW_i * BOA_i$	1.405 (3.177)	1.105 (2.288)
$SIZE_i$	0.491 (11.166)	0.415 (8.654)
LEV_i	0.004 (0.091)	0.031 (0.670)
SEG_i	0.114 (4.095)	0.085 (2.775)
SOE_i	0.120 (1.936)	0.124 (1.805)
OWN_i	0.003 (2.293)	0.003 (1.619)
$NAUDIT_i$	0.124 (1.797)	0.189 (2.485)
$BIG4_i$	0.154 (1.642)	0.313 (3.068)
$INDI_i$	−0.010 (−1.837)	−0.010 (−1.713)
$LITI_i$	0.083 (1.471)	0.109 (1.754)
$Industry_i$	Control	Control
$Adj \cdot R^2$	0.337	0.276
F value	34.060	25.447

We think occurrence of the above conclusions is because that *corporate internal control guidance* is implemented just now and related evaluation, disclosure and supervision systems are not perfect enough. Therefore, companies may have internal control deficiencies possibly choose to formalize its internal control deficiency or conceal its internal control deficiency in order to prevent bringing negative influence after deficiency disclosure. However, enterprises with higher executive compensation level in China mostly have characteristics such as good capital quality; pay more attention to public interest entity image and so on. As the above enterprises receive attention from more aspects, such as minority shareholder, media and supervision departments and so on, management need to measure its internal control level accurately and disclose it to public constrained by external pressure. At the same time, due to existence of the above phenomena, disclosing its internal control deficiency not only will not harm corporate market images, instead it will transmit a forward signal that corporate management can evaluate its internal control objectively and fairly, input enough resources to its internal control building and the internal control system is in constant perfect stage to the market. Therefore, different with existing research results based on American capital market, disclosing internal control deficiency by familiar listed companies in China presents an inverted phenomenon, namely familiar listed companies with higher quality are more inclined to disclose its existing deficiency of internal control and Board of Directors and management will be more inclined to disclose its existing deficiency of internal control and disclose related rectification information driven by high-level compensation. Group samples are adopted for further verification of studied problem in the Thesis for deep exploration.

We implemented Chow test on group samples and Chow test result of three groups of sub sample are significant at

the level of more than 10%, therefore, group standard selection in the Thesis has reasonability. Based on group test result of disclosing internal control deficiency, we noticed that executive compensation has significant performance sensibility in groups of disclosing internal control deficiency, but performance sensibility of executive compensation is not significant in groups without disclosing internal control deficiency. At the same time, we noticed that executive compensation is significantly and positively related to corporate financial leverage in internal control deficiency group, while executive compensation is significantly and negatively related to corporate financial leverage in groups without disclosing deficiencies. By analyzing the reasons, we think that though corporate liabilities can reduce overall capital cost of company and thus improve corporate business performance, financial risk of company increase together and higher business pressure will be caused, so it is easier to expose existing deficiency of corporate internal control. At the same time, based on agency theory, creditor has stronger supervision motivation compared with shareholders. Therefore, with improvement of corporate financial leverage, and enhancement of supervision motivation of creditor, Board of Directors and management will make more accurate evaluation for internal control system and disclose existing problems of internal control. Meanwhile, we found that frequency of independent directors performing responsibilities is negatively related to executive compensation at the level of 5% in deficiency disclosing group, which indicates that independent director system in China plays certain supervision role in executive behaviors. Based on performance level group, being similar with test result of complete sample, group test result shows that executive compensation is positively related to internal control.

5 Conclusions

Executive compensation level of familiar listed company in China is significantly and negatively related to internal control deficiency at present stage, namely those enterprises with higher executive compensation level are more inclined to disclose its internal control deficiency, because internal control deficiency information disclosed by familiar listed company in China is to transmit favorable new outward, and have investor and social public decision value. After implementation of *corporate internal control guidance,* internal control deficiency disclosure improves performance sensitivity of executive compensation significantly. By further test of group sample, based on performance level group, we found that significance of positive relation between internal control deficiency disclosure of familiar listed companies and executive compensation level in groups with higher performance level is higher than that of companies with lower performance level; however internal control deficiency disclosure will increase its performance sensitivity of executive compensation significantly in groups with lower performance level; in group test of ownership property, we found that disclosing internal control deficiency can increase executive performance compensation sensitivity in state-owned enterprises; executive compensation level in non state-owned enterprises is significantly and negatively related to internal control deficiency.

According to research conclusions, the following suggestions are proposed: internal control deficiency disclosure of familiar listed companies in China has inverted phenomenon on the whole at present stage. Therefore, it is suggested for supervision departments to strengthen supervision strength and improve effectiveness of internal control evaluation and its deficient disclosure in familiar listed companies; in the aspect of enterprise, evaluation mechanism of executive compensation shall not only be based on performance and other financial indexes, but also take non-financial indexes with higher information contents into consideration at the same time, further perfect making compensation incentive mechanism of familiar listed companies and improve the whole value of company.

Acknowledgements The Natural Science Foundation of Kunming University of Science and Technology (provincial talent training project), No. KKSY201508012, and Science Fund Projects of Yunnan Provincial Education Department, No. 2017ZZX160.

References

1. Kim, Y., Park, M.S.: Market uncertainty and disclosure of internal control deficiencies under the Sarbanes–Oxley Act. J. Acc. Public Policy **28**(5), 419–445 (2009)
2. Zhao, X., Ning-Ning, X.U., Management, S.O., et al.: Element analysis of management's internal control deficiencies disclosure activity—based on the view of the motivated selectivity in the internal control information disclosure. Econ. Surv. **31**(3), 85–89 (2014)
3. Ji, X.D., Lu, W., Qu, W.: Determinants and economic consequences of voluntary disclosure of internal control weaknesses in China. J. Contemp. Acc. Econ. **11**(1), 1–17 (2015)
4. Wang, X., Gao, J., Lu, Q.: A literature review of internal control deficiencies. In: Proceedings of the international conference on service systems and service management, pp. 1–4. IEEE (2015)
5. Hamza, R., Muhammad, K., Arunkumar, N, Ramírez González, G.: Hash based encryption for keyframes of diagnostic hysteroscopy. IEEE Access (2017). https://doi.org/10.1109/ACCESS.2017.2762405
6. Fernandes, S.L., Gurupur, V.P., Sunder, N.R., Arunkumar, N., Kadry, S. A novel nonintrusive decision support approach for heart rate measurement. Pattern Recognit. Lett. (2017). https://doi.org/10.1016/j.patrec.2017.07.002
7. Arunkumar, N., Ramkumar, K., Venkatraman, V., Abdulhay, E., Fernandes, S.L., Kadry, S., Segal, S.: Classification of focal and non focal EEG using entropies. Pattern Recognit. Lett. **94**, 112–117 (2017)

Wendai Lv received her doctor's degree in technological economics and management from Renming University of China. She is currently a lecturer in the College of Management and Economics, Kunming University of Science and Technology. Her research interest is mainly in the area of Internal control, Optimizing Decision, Risk Management, Corporate Governance. She has published several research papers in scholarly journals in the above research areas and has participated in several books.